Database resources of the National Center for Biotechnology Information: 2002 update

David L. Wheeler*, Deanna M. Church, Alex E. Lash, Detlef D. Leipe, Thomas L. Madden, Joan U. Pontius, Gregory D. Schuler, Lynn M. Schriml, Tatiana A. Tatusova, Lukas Wagner and Barbara A. Rapp

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health,
Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 20, 2001; Accepted September 26, 2001

ABSTRACT

In addition to maintaining the GenBank nucleic acid sequence database, the National Center for Biotechnology Information (NCBI) provides data analysis and retrieval resources that operate on the data in GenBank and a variety of other biological data made available through NCBI's web site. NCBI data retrieval resources include Entrez, PubMed, LocusLink and the Taxonomy Browser. Data analysis resources include BLAST, Electronic PCR, OrfFinder, RefSeq, UniGene, HomoloGene, Database of Single Nucleotide Polymorphisms (dbSNP), Human Genome Sequencing, Human MapViewer, Human¡VMouse Homology Map, Cancer Chromosome Aberration Project (CCAP), Entrez Genomes, Clusters of Orthologous Groups (COGs) database, Retroviral Genotyping Tools, SAGEmap, Gene Expression Omnibus (GEO), Online Mendelian Inheritance in Man (OMIM), the Molecular Modeling Database (MMDB) and the Conserved Domain Database (CDD). Augmenting many of the web applications are custom implementations of the BLAST program optimized to search specialized data sets. All of the resources can be accessed through the NCBI home page at http://www.ncbi.nlm.nih.gov.

INTRODUCTION

The National Center for Biotechnology Information (NCBI) at the National Institutes of Health was created in 1988 to develop information systems for molecular biology. In addition to maintaining the GenBank (1) nucleic acid sequence database, to which data is submitted directly by the scientific community, NCBI provides data retrieval systems and computational resources for the analysis of GenBank data and the variety of other biological data made available through NCBI. For this update, the NCBI suite of database resources is grouped into the seven categories given below. All resources discussed below are available from the NCBI home page at http://www.ncbi.nlm.nih.gov.

DATABASE RETRIEVAL TOOLS

Entrez

Entrez (2) is an integrated database retrieval system that accesses DNA and protein sequences derived from several sources (1,3–6), genome maps, population sets, protein structures from the Molecular Modeling Database (MMDB) (7) and the biomedical literature via PubMed and Online Mendelian Inheritance in Man (OMIM), with embedded links to the NCBI taxonomy. PubMed includes primarily the 11 million references and abstracts in MEDLINE, with links to the full-text of more than 1100 journals available on the web.

Several new features have been added to Entrez recently including searches for proteins by molecular weight range or by protein domain. Population sets and 'ProbeSet' data from gene-expression experiments are also now accessible from Entrez. New Entrez formatting options include graphical and text-mode display of PHRAP scores for HTG records and XML output for all GenBank records.

The Books database has been integrated into Entrez allowing the searching of three online scientific textbooks (8–10).

LocusLink

The LocusLink database of official gene names and other gene identifiers (6), was developed at NCBI in conjunction with several international collaborators, and offers a single query interface to curated sequences and descriptive information about genes. New links from the LocusLink report lead to Proteome Inc.'s Gene Ontology, and NCBI's Evidence Viewer which shows the alignments between GenBank sequences and the NCBI Human Genome Assembly used to derive alignment-based gene models. Protein domains from NCBI's Conserved Domain Database (CDD) that are detected in gene translation products are also now listed in the LocusLink report.

THE BLAST FAMILY OF SEQUENCE-SIMILARITY SEARCH PROGRAMS

The Basic Local Alignment Search Tool (BLAST) family of programs (11,12) performs sequence-similarity searches, beginning with either a query sequence or a GenBank accession number. Successful searches return a set of gapped alignments between the query and similar database sequences, with links

*To whom correspondence should be addressed. Tel: +1 301 435 5950; Fax: +1 301 480 9241; Email: wheeler@ncbi.nlm.nih.gov to the full database records. Each alignment receives a score and a measure of statistical significance, called the Expectation Value, for judging its quality.

The NCBI BLAST interface has been re-designed and offers several new search options including the specification of an Expectation Value range, rather than a threshold, for reporting alignments, and the specification of a residue range to limit searches to a portion of the query sequence. XML output is now supported. A new alignment format, called the 'Hit Table', provides a compact, tabular summary of the BLAST results including, for each database hit, the positions of alignment starts and stops, coupled with scores and Expectation Values. In addition, BLAST can generate a taxonomically organized output that shows the distribution of BLAST hits by organism in three formats.

A particularly powerful feature of the new BLAST interface allows searches to be restricted to a database subset using standard Entrez search strings; the same restrictions may be applied to screen the output of an initially unrestricted search. These features provide the means to effectively construct a custom database for searching, or to parse the output of a search to include only sequences of interest, respectively.

Along with the revised BLAST interface, NCBI has implemented a standard URL-API which allows complete search specifications, including BLAST parameters and search query, to be contained in the URL posted to the web page. A 'GetURL' button on the BLAST pages allows for the saving of the current parameter set, but URLs for custom searches may also be constructed easily by users.

On the algorithmic side, BLAST now takes into account the amino acid composition of the query sequence in its estimation of statistical significance. A composition-based statistical treatment, used in conventional protein BLAST searches as well as PSI-BLAST (12) searches, tends to reduce the number of false-positive database hits (13).

A useful BLAST utility, BLAST2Sequences (14), compares two DNA or protein sequences and produces a dot-plot representation of the alignments it reports. Translated search options, such as blastx, tblastx and tblastn, now extend the the program's range beyond blastn and blastp.

Using a new nucleotide BLAST variant, called MegaBLAST (15), batches of nucleotide sequences can be pasted into a web page or uploaded from a file, and used to search for nearly exact matches in nucleotide databases. MegaBLAST is up to 10 times faster than BLASTn for such searches.

NCBI has developed a semi-automated system for assembling both finished and unfinished human genome sequence on a regular basis so as to incorporate the most current data. The NCBI-generated assembly of the human genome may be searched via a specialized variant called Human Genome BLAST using either nucleotide or protein queries. Human Genome BLAST generates custom 'Genome view' of the BLAST hits which is integrated with the Human Genome MapViewer so that the hits can be viewed in the context of a combination of the maps used by the MapViewer, such as maps showing confirmed and predicted gene locations, or EST hits.

Finally, a special database, called the Trace Archive, which contains raw data underlying sequences generated by the various genome projects, may be searched using MegaBLAST. The Trace Archive contains Whole Genome Shotgun (WGS) reads from the mouse as well as data from rat, human, zebrafish and worm.

Blink

BLAST-Link (Blink) is a new resource that displays pre-computed protein BLAST alignments for each protein sequence in the Entrez databases. Blink allows for the display of subsets of these alignments by taxonomic criteria, by database of origin, relation to a complete genome, membership in a COG (16) or by relation to a 3D structure or conserved protein domain. Blink links are displayed for protein records in Entrez as well as within LocusLink reports.

RESOURCES FOR GENE-LEVEL SEQUENCES

UniGene

UniGene (17) is a system for automatically partitioning GenBank mRNA sequences and EST, into a non-redundant set of gene-oriented clusters with identical 3′ untranslated regions (3′ UTRs). Six new organisms have been added to UniGene over the last year: barley, rice, Thale cress, maize, frog and wheat. These organisms join human, mouse, rat, zebrafish and cow, bringing the total number of organisms represented to 11. The UniGene databases are updated weekly with new EST sequences, and bimonthly with newly characterized sequences and may be downloaded by FTP.

HomoloGene

HomoloGene is a database of both curated and calculated gene orthologs and homologs for the human, mouse, rat, zebrafish, cow, frog and fly; the latter two organisms added over the last year. There are now over 50 000 homology groups for human and mouse. The current datasets for the calculated orthologs and homologs and Mutually Orthologous Pairs are available via FTP.

RefSeq

The References Sequence (RefSeq) database (6), provides curated reference sequences for mRNAs and proteins from human and other organisms. A new type of RefSeq, with accession number of the form NG_######, has been created to accommodate curated genomic regions. Such regions now include the human hair keratin gene cluster on chromosome 17, the human α globin region on chromosome 16, and the human MHC class III complement gene cluster on chromosome 6, among others.

dbSNP

The database of Single Nucleotide Polymorphisms (dbSNP) (18) serves as a repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms and now contains almost 3 million human SNPs. A new web interface allows flexible searches by gene name and by cross-reference to other databases such as OMIM or the structure databases. Searches for SNPs lying between two markers and batch retrieval are now also supported.

e-PCR

Electronic PCR (e-PCR) is a tool for locating STSs within a nucleotide sequence by comparing the query against a database of STS sequences and primer pairs. A new non-redundant database of over 133 000 human and 80 000 non-human STSs called UniSTS is now available for such comparisons.

RESOURCES FOR CHROMOSOMAL SEQUENCES

The Human Genome MapViewer, Human Genome Assembly and Human Genome Resources

The Human Genome MapViewer displays NCBI's current human genome assembly using up to seven parallel chromosomal maps simultaneously. The maps displayed can be selected from a set of 21, and include cytogenetic maps, such as chromosomal ideograms, sequence-based maps, such as those showing contigs, genes and SNPs, and radiation hybrid maps, such as the G3 and GB4 maps. Several new maps have recently been added to show *ab initio* gene models predicted by GenomeScan, EST alignments with links to UniGene clusters and mRNA alignments used to construct gene models. The MapViewer also offers a tabular view of the data, in addition to the default graphical view, which is convenient for the import of the data into other programs for further analysis, and can be used to download user-selected segments of the NCBI Human Genome Assembly using a 'Download/View Sequence' link. Supported download formats are GenBank and FASTA. The NCBI Human Genome Assembly data are also available via FTP.

The Human Genome Resources web page collects links to all of NCBI's human genome-related tools on one page. A recent addition to these links is the Human BAC Resource, a catalog of large-insert, FISH-mapped clones containing sequence-tagged sites that can be used to integrate cytogenetic, radiation-hybrid, linkage and sequence maps of the human genome and to locate clone distributors.

The Cancer Chromosome Aberration Project (CCAP)

The CCAP service is an initiative of the National Cancer Institute (NCI) and NCBI. The data includes a compilation by F. Mitelman, F. Mertens and B. Johansson of recurrent neoplasia-associated chromosomal aberrations from the Cancer Chromosome Aberration Bank at the University of Lund, Sweden (19). A new Spectral Karyotyping database, SKY, has been created jointly by NCI and NCBI to enable investigators to share their own SKY and Comparative Genomic Hybidization (CGH) data on chromsomal aberrations (http://www.ncbi.nlm.nih.gov/sky/skyweb.cgi).

RESOURCES FOR GENOME-SCALE ANALYSIS

Entrez Genomes

Entrez Genomes (20) provides access to genomic data contributed by the scientific community for over 900 species whose sequencing and mapping is complete or in progress, and now includes more than 57 complete microbial genomes and 245 reference sequences for eukaryotic organelles; a gain of about 27 microbial genomes and 80 organelle sequences over last year. Four higher eukaryotic genomes are also found within Entrez Genomes; *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Arabidopsis thaliana*. A new tool for genomic comparisons has been added within the last year called TaxPlot, which graphically compares similarities in the proteomes of two organisms to that of a third reference organism.

Entrez Genomes reports for complete microbial genomes give access to views showing various aspects of proteome organization , including the taxonomic distribution of proteins encoded in the genome (TaxMaps) and their membership in Clusters of Orthologous Groups (COGs) (16), links to 3D structures and conserved protein domains. Pairwise sequence alignments for sequence neighbors are presented graphically and linked to the Cn3D macromolecular viewer (21), which allows the interactive display of 3D structures and sequence alignments.

COGs

The COGs database (16) presents a compilation of orthologous groups of proteins from completely sequenced organisms representing 30 phylogenetically distant clades. Many new organisms have been added over the last year, bringing the total to over 44. The COGs are now also linked to the proteins of two higher eukaryotes; the worm and the fly.

RESOURCES FOR THE ANALYSIS OF PATTERNS OF GENE EXPRESSION AND PHENOTYPES

SAGEmap

NCBI's SAGEmap service implements many functions useful in the analysis of SAGE data such as a two-way mapping between SAGE tag and UniGene. A new Java-Based SAGEmap Submission Tool (SST) is now available to assist SAGEmap submissions, and may be useful as a SAGE data organizational and analysis tool.

OMIM

NCBI provides web access to the OMIM catalog of human genes and genetic disorders authored and edited by Victor A. McKusick at The Johns Hopkins University (22). The database contains information on disease phenotypes and genes, including extensive descriptions, gene names, inheritance patterns, map locations and gene polymorphisms. OMIM currently contains 12 940 entries, including data on 9526 established gene loci and 926 phenotypic descriptions, and is now searchable using the powerful Entrez interface.

GEO

The Gene Expression Omnibus (GEO), a data repository and retrieval system for gene expression data, is described in a separate article in this issue (23).

THE MMDB, THE CONSERVED DOMAIN DATABASE SEARCH AND DART

The NCBI MMDB, built by processing entries from the Protein Data Bank (5), is described by Wang *et al.* (7). The structures in the MMDB are linked to sequences in Entrez and to the CDD. The CDD contains PSI-BLAST-derived Position Specific Score Matrices (PSSM's) representing domains derived principally from two public protein domain collections, the Simple Modular Architecture Research Tool (SMART) (24) and Pfam (25), but also draws from domains defined by NCBI researchers. NCBI's Conserved Domain Search (CD-Search) service can be used to search a protein sequence for conserved domains in the CDD. Wherever possible CDD hits are linked to structures which, coupled with a multiple sequence alignment of representatives of the domain hit, can be viewed with NCBI's 3D molecular structure viewer, Cn3D (21). A new Tool, called the Domain Architecture Retrieval Tool (DART), has been added to the MMDB suite which allows searches of protein databases on the basis of a combination of conserved domains. Protein domain information has also been integrated into the Entrez system as mentioned above.

FOR FURTHER INFORMATION

Most of the resources described here include documentation, other explanatory material and references to collaborators and data sources on the respective web sites. Several tutorials are also offered under the Education link from NCBI's home page. A Site Map provides a comprehensive table of NCBI resources, and the What's New feature announces new and enhanced resources. A user support staff is available to answer questions at info@ncbi.nlm.nih.gov.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 17–20.
2. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
3. Barker,W.C., Garavelli,J.S., Huong,H., McGarvey,P.B., Orcutt,B.C., Srinivarsarao,G.Y., Xiao,C., Yeh,L.S., Ledley,R.S., Janda,J., Pfeiffer,F., Mewes,H.W., Tsugita,A. and Wu,K. (2000) The Protein Information Resource (PIR). *Nucleic Acids Res.*, **28**, 41–44. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 35–37.
4. Kriventseva,E.V., Fleischmann,W., Zdobnov,E.M. and Apweiler,R. (2001) CluSTr: a database of Clusters of SWISS-PROT and TrEMBL proteins. *Nucleic Acids Res.*, **29**, 33–36.
5. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 245–248.
6. Pruitt,K. and Maglott,D. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
7. Wang,Y., Addess,K.J., Geer,L., Madej,T., Marchler-Bauer,A., Zimmerman,D. and Bryant,S.H. (2000) MMDB: 3D structure data in Entrez. *Nucleic Acids Res.*, **28**, 243–245. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 249–252.
8. Riddle,D.L., Blumenthal,T., Meyer,B.J. and Priess,J.R. (1997) *C. elegans II.* Cold Spring Harbor Laboratory Press, Plainview, NY.
9. Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J.D. (1994) *Molecular Biology of the Cell*, 3rd Edn. Garland Publishing, New York, London.
10. Coffin,J.M., Hughes,S.H. and Varmus,H.E. (1997) *Retroviruses.* Cold Spring Harbor Laboratory Press, Plainview, NY.
11. Altschul,S.E., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
13. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
14. Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, **174**, 247–250.
15. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comp. Biol.*, **7**, 203–214.
16. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
17. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
18. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Pham,L., Smigielski,E. and Sirotkin,K. (2001) dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
19. Mitelman,F., Mertens,F. and Johansson,B. (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genet.*, **15**, 417–474.
20. Tatusova,T., Karsch-Mizrachi,I. and Ostell,J. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.
21. Wang,Y., Geer,L.Y., Chappey,C., Kans,J.A. and Bryant,S.H. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem. Sci.*, **25**, 300–302.
22. McKusick,V.A. (1998) *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders*, 12th Edn. The Johns Hopkins University Press, Baltimore, MD.
23. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
24. Schultz,J., Copley,R.R., Doerks,T., Ponting,C.P. and Bork,P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, **28**, 231–234. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 242–244.
25. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 276–280.